On Fast Convergence of Proximal Algorithms for SQRT-Lasso Optimization: DonâĂŹt Worry About its Nonsmooth Loss Function

Xingguo Li[◇], Haoming Jiang[†], Jarvis Haupt[*], Raman Arora[⋆], Han Liu[○], Mingyi Hong[⋆], Tuo Zhao[†]
[◇]Princeton University [*]University of Minnesota [⋆]Johns Hopkins University [○]Northwestern University [†]Georgia Tech

Background

Linear Model: $y = X\theta + \epsilon$, with $\mathbb{E}[\epsilon] = 0$ and $\mathbb{E}[\epsilon^2] = \sigma^2$
SQRT(square-root) Lasso regression:

$$\overline{\theta}^{\mathsf{SQRT}} = \mathrm{argmin}_{\theta\in\mathbb{R}^d} \mathcal{F}_\lambda(\theta) = \underbrace{\frac{1}{\sqrt{n}}||y - X\theta||_2}_{\text{Loss } \mathcal{L}(\theta)} + \underbrace{\lambda_{\mathsf{SQRT}}||\theta||_1}_{\text{Regularizer}}, \quad (1)$$

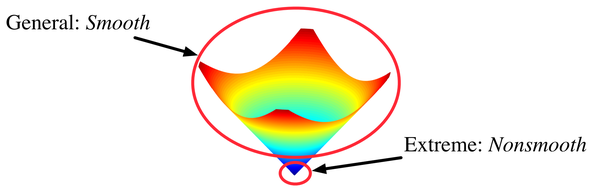
- $\lambda_{\mathsf{SQRT}} > 0$: the regularization parameter.

Lasso vs. **SQRT-Lasso**:

$$\overline{\theta}^{\mathsf{Lasso}} = \mathrm{argmin}_\theta \tfrac{1}{n}||y - X\theta||_2^2 + \lambda_{\mathsf{Lasso}}||\theta||_1,$$

- Lasso: **smooth** loss, **tuning heavy** $\lambda_{\mathsf{Lasso}} \asymp \sigma\sqrt{\frac{\log d}{n}}$
- SQRT-Lasso: **nonsmooth** loss, **tuning free** $\lambda_{\mathsf{SQRT}} \asymp \sqrt{\frac{\log d}{n}}$

Motivation: "almost" smooth loss function. The ℓ_2 loss function is not differentiable only when the model are overfitted,



Proximal Algorithms

Prox-Grad: Given $\theta^{(t)}$ at t-th iteration, we consider minimizing a quadratic approximation of (1) as
$\theta^{(t+1)} = \mathcal{S}_{\frac{\lambda}{L^{(t)}}}\left(\theta^{(t)} - \frac{\nabla\mathcal{L}(\theta^{(t)})}{L^{(t)}}\right) = \mathrm{argmin}_\theta \mathcal{Q}_\lambda(\theta, \theta^{(t)})$,
where $\mathcal{Q}_\lambda(\theta, \theta^{(t)}) = \mathcal{L}(\theta^{(t)}) + \nabla\mathcal{L}(\theta^{(t)})^\top(\theta - \theta^{(t)}) + \frac{L^{(t)}}{2}||\theta - \theta^{(t)}||_2^2 + \lambda||\theta||_1$, $L^{(t)}$ is step size determined by line search.

Algorithm 1 Proximal Gradient Algorithm

Input: y, X, λ, ε, $L_{\max} > 0$; **Initialize:** $\theta^{(0)}$, $t \leftarrow 0$;
Repeat: $t \leftarrow t + 1$
$L^{(t)} \leftarrow$ Line Search
$\theta^{(t)} \leftarrow \mathcal{S}_{\lambda/L^{(t)}}\left(\theta^{(t-1)} - \nabla\mathcal{L}(\theta^{(t-1)})/L^{(t)}\right)$
Until: $\omega_\lambda(\theta^{(t)}) \le \varepsilon$; **Return:** $\widehat{\theta} \leftarrow \theta^{(t)}$

Prox-Newton: Given $\theta^{(t)}$ at t-th iteration, we instead update θ as
$\theta^{(t+0.5)} = \mathrm{argmin}_\theta \mathcal{Q}_\lambda(\theta, \theta^{(t)})$ and
$\theta^{(t+1)} = \theta^{(t)} + \eta_t(\theta^{(t+0.5)} - \theta^{(t)})$ by line search,
where $\mathcal{Q}_\lambda(\theta, \theta^{(t)}) = \mathcal{L}(\theta^{(t)}) + \nabla\mathcal{L}(\theta^{(t)})^\top(\theta - \theta^{(t)}) + \frac{1}{2}(\theta - \theta^{(t)})^\top\nabla^2\mathcal{L}(\theta^{(t)})(\theta - \theta^{(t)}) + \lambda||\theta||_1$.

Algorithm2 Proximal Newton Algorithm

Input: y, X, λ, ε; **Initialize:** $\theta^{(0)}$, $t \leftarrow 0$;
Repeat: $t \leftarrow t + 1$;
$\theta^{(t)} \leftarrow \mathrm{argmin}_\theta \mathcal{Q}_\lambda(\theta, \theta^{(t-1)})$
$\eta_t \leftarrow$ Line Search
$\theta^{(t)} \leftarrow \theta^{(t)} + \eta_t(\theta^{(t)} - \theta^{(t-1)})$
Until: $\omega_\lambda(\theta^{(t)}) \le \varepsilon$; **Return:** $\widehat{\theta} \leftarrow \theta^{(t)}$

Stopping Criteria (*approximate KKT condition*):
$\omega_\lambda(\theta^{(t)}) = \min_{g\in\partial||\theta^{(t)}||_1} ||\nabla\mathcal{L}(\theta^{(t)}) + \lambda g||_\infty \le \varepsilon$.

Geometric Properties of SQRT-Lasso

Lemma 1. *Given sub-Gaussian assumpotion of X and ϵ, and $||\theta^*||_0 = s^*$, then with high probability:*
(i) *for $\forall\lambda \ge C_1\sqrt{\frac{\log d}{n}}$, we have*

$$\lambda \ge \frac{C_1}{4}||\nabla\mathcal{L}(\theta^*)||_\infty,$$

which is sufficiently large to yield sufficiently sparse solutions;

(ii) *for $n \ge C_2 s^* \log d$, $\exists\mathcal{B}_r = \{\theta\in\mathbb{R}^d : ||\theta - \theta^*||_2^2 \le r\}$ for some constant r s.t. for $\forall v, w \in \mathcal{B}_r$ satisfying $||v - w||_0 \le s^* + 2\widetilde{s}$, $\mathcal{L}$ is locally restricted strongly convex (LRSC), smooth (LRSS), and Hessian smooth (LRHS), if $\exists\rho_s^-, \rho_s^+, L_s \in (0,\infty)$ i.e.*

$$\textit{LRSC:}\mathcal{L}(v) - \mathcal{L}(w) - \nabla\mathcal{L}(w)^\top(v - w) \ge \frac{\rho_s^-}{2}||v - w||_2^2,$$

$$\textit{LRSS:}\mathcal{L}(v) - \mathcal{L}(w) - \nabla\mathcal{L}(w)^\top(v - w) \le \frac{\rho_s^+}{2}||v - w||_2^2,$$

$$\textit{LRHS:} \min_{\substack{||u||_0\le s,\\ ||u||_2=1}} u^\top(\nabla^2\mathcal{L}(v) - \nabla^2\mathcal{L}(w))u \le L_s||v - w||_2^2,$$

with

$$\rho_{s^*+2\widetilde{s}}^+ \le \frac{C_3}{\sigma},\ \rho_{s^*+2\widetilde{s}}^- \ge \frac{C_4}{\sigma} \text{ and } L_{s^*+2\widetilde{s}} \le \frac{C_5}{\sigma},$$

where $C_1, \ldots, C_5 \in \mathbb{R}^+$ are generic constants.

Computational Theory

Denote

$$\kappa_s = \frac{\rho_s^+}{\rho_s^-}, \quad \mathcal{S}^* = \{j \mid \theta_j^* \ne 0\}, \quad \overline{\mathcal{S}}^* = \{j \mid \theta_j^* = 0\},$$
$$\mathcal{B}_r^{s^*+\widetilde{s}} = \mathcal{B}_r \cap \{\theta\in\mathbb{R}^d : ||\theta - \theta^*||_0 \le s^* + \widetilde{s}\}.$$

Theorem 2 (Local Convergence). *Suppose X and n satisfy Lemma 1. Given λ and $\theta^{(0)}$ such that $\lambda \ge \frac{C_1}{4}||\nabla\mathcal{L}(\theta^*)||_\infty$, $||\theta^{(0)} - \theta^*||_2^2 \le s^*\left(8\lambda/\rho_{s^*+\widetilde{s}}^-\right)^2$ and $\theta^{(0)} \in \mathcal{B}_r^{s^*+\widetilde{s}}$, we have sufficiently sparse solutions for all iterations, i.e.,*

$$||[\theta^{(t)}]_{\overline{\mathcal{S}}^*}||_0 \le \widetilde{s}.$$

Moreover, given $\varepsilon > 0$, we need at most

$$T = \mathcal{O}\left(\kappa_{s^*+2\widetilde{s}}\log\left(\frac{\kappa_{s^*+2\widetilde{s}}^3 s^*\lambda^2}{\varepsilon^2}\right)\right) \quad \textit{Prox-Grad}$$

$$T = \mathcal{O}\left(\log\log\left(\frac{3\rho_{s^*+2\widetilde{s}}^+}{\varepsilon}\right)\right) \quad \textit{Prox-Newton}$$

iterations to guarantee that the output solution $\widehat{\theta}$ satisfies

$$||\widehat{\theta} - \overline{\theta}||_2^2 = \mathcal{O}\left(\left(1 - \frac{1}{8\kappa_{s^*+2\widetilde{s}}}\right)^T \varepsilon\lambda s^*\right) \quad \textit{Prox-Grad}$$

$$||\widehat{\theta} - \overline{\theta}||_2^2 = \mathcal{O}\left(\left(\frac{L_{s^*+2\widetilde{s}}}{2\rho_{s^*+2\widetilde{s}}^-}\right)^{2^T} \varepsilon\lambda s^*\right) \quad \textit{Prox-Newton,}$$

where $\overline{\theta}$ is the unique sparse global optimum to (1) with $||[\overline{\theta}]_{\overline{\mathcal{S}}^*}||_0 \le \widetilde{s}$.

Statistical Theory

Theorem 3. *Suppose X, and n satisfy conditions in Lemma 1. Given $\lambda = C_1\sqrt{\log d/n}$, if the output solution $\widehat{\theta}$ obtained from Algorithm 1 and 2 satisfying, $\omega_\lambda(\widehat{\theta}) \le \varepsilon = \mathcal{O}\left(\frac{\sigma s^*\log d}{n}\right)$, then we have:*

$$||\widehat{\theta} - \theta^*||_2 = \mathcal{O}_P\left(\sigma\sqrt{\frac{s^*\log d}{n}}\right) \quad \textit{and}$$
$$||\widehat{\theta} - \theta^*||_1 = \mathcal{O}_P\left(\sigma s^*\sqrt{\frac{\log d}{n}}\right).$$

Moreover, we have $|\widehat{\sigma} - \sigma| = \mathcal{O}_P\left(\frac{\sigma s^*\log d}{n}\right)$, where $\widehat{\sigma} = \frac{||y - X\widehat{\theta}||_2}{\sqrt{n}}$.

In Algorithm 1,2, $\widehat{\theta}$ achieves **the minimax optimal rate of convergence** in parameter estimation.

Boosting via Pathwise Optimization Scheme

Algorithm 3 The pathwise optimization scheme

Input: y, X, N, $\lambda_{[N]}$, $\varepsilon_{[N]}$
Initialize: $\widehat{\theta}_{[0]} \leftarrow 0$, $\lambda_{[0]} \leftarrow ||\nabla\mathcal{L}(0)||_\infty$, $\eta_\lambda \leftarrow \left(\frac{\lambda_{[N]}}{\lambda_{[0]}}\right)^{\frac{1}{N}}$
For: $K = 1, \ldots, N$
$\lambda_{[K]} \leftarrow \eta_\lambda\lambda_{[K-1]}$, $\theta_{[K]}^{(0)} \leftarrow \widehat{\theta}_{[K-1]}$, $\varepsilon_{[K]} \leftarrow \varepsilon_{[N]}$
$\widehat{\theta}_{[K]} \leftarrow$ Prox-Alg $\left(y, X, \lambda_{[K]}, \theta_{[K]}^{(0)}, \varepsilon_{[K]}\right)$
End For
Return: $\widehat{\theta}_{[N]}$

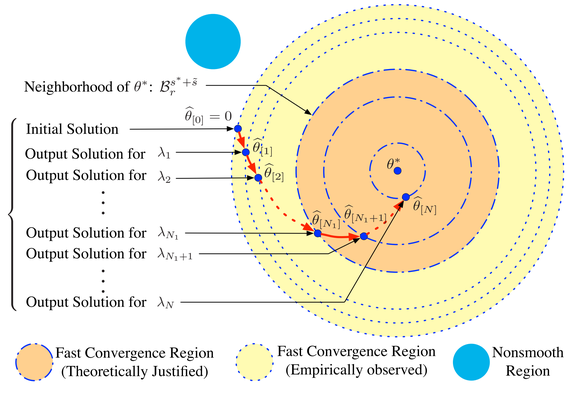
Theorem 4. *Suppose the design matrix X is sub-Gaussian, and $\lambda_{[N]} = C_1\sqrt{\log d/n}$. For $n \ge C_2 s^*\log d$ and $\eta_\lambda \in (\frac{5}{6}, 1)$, the following results hold:*
(I) *There exists an $N_1 < N$ such that*

$$r > s^*\left(8\lambda_{N_1}/\rho_{s^*+\widetilde{s}}^-\right)^2;$$

(II) *For any $K \in [N_1 + 1, .., N]$, we have*

$$||\theta_{[K]}^{(0)} - \theta^*||_2^2 \le s^*\left(8\lambda_{[K]}/\rho_{s^*+\widetilde{s}}^-\right)^2, \theta_{[K]}^{(0)} \in \mathcal{B}_r^{s^*+\widetilde{s}};$$

(III) *Theorems 2 hold for all λ_K's, where $K \in [N_1 + 1, .., N]$.*



Experiments

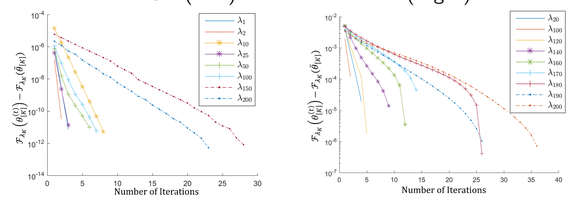
Datasets:

- "Greenhouse" ($n = 2921$, $d = 5232$, [2])
- "DrivFace" ($n = 606$, $d = 6400$, [1])
- simulated datasets: ($n = 1000$, $d =$1k, 5k, 10k)

Tab. 1 Computational performance of Prox-GD on synthetic data under different choices of variance σ, the number of stages N, and the stopping criterion ε_N.

σ	N	ε_N 10^{-4}	10^{-5}	10^{-6}	Minimal MSE
0.1	1	0.372	0.372	0.365	0.013
	10	**0.275**	**0.276**	**0.280**	
	30	0.336	0.345	0.351	
1	1	0.235	0.248	0.262	1.183
	10	**0.104**	**0.103**	**0.109**	
	30	0.217	0.222	0.220	

Fig. 1 The objective gap v.s. the number of iterations for Prox-GD (Left) and Prox-Newton (Right).



Tab. 2 Timing comparison between algorithms on real data.

Data Set	SQRT-Lasso						Lasso
	Prox-GD	Newton	ADMM	ScalReg	CD	Alt.Min	PISTA
Greenhouse	5.812	**1.708**	1027	3181	14.31	99.81	5.113
DrivFace	**0.421**	0.426	18.88	124.0	3.138	17.69	0.414

Tab. 3 Extension to calibrated multivariate regression.

λ_N	Synthetic ($\sigma = 1$)			
	Prox-GD	Newton	ADMM	CD
$\sqrt{\log d/n}$	0.2964	**0.0320**	14.83	2.410
$2\sqrt{\log d/n}$	0.1725	**0.0213**	2.231	2.227
$4\sqrt{\log d/n}$	0.0478	**0.0112**	1.868	1.366
λ_N	DrivFace			
$\sqrt{\log d/n}$	9.562	**0.2186**	158.9	12.77
$2\sqrt{\log d/n}$	8.688	**0.1603**	129.4	20.42
$4\sqrt{\log d/n}$	1.824	**0.0924**	94.37	19.17

References

[1] K. Diaz-Chito, A. Hernández-Sabaté, and A. M. López. A reduced feature set for driver head pose estimation. *Appl. Soft Comput.*, 45(C):98–107, Aug. 2016.

[2] D. D. Lucas, C. Yver Kwok, P. Cameron-Smith, H. Graven, D. Bergmann, T. P. Guilderson, R. Weiss, and R. Keeling. Designing optimal greenhouse gas observing networks that consider performance and cost. *Geoscientific Instrumentation, Methods and Data Systems*, 4(1):121–137, 2015.

Acknowledgment

PRINCETON UNIVERSITY, UNIVERSITY OF MINNESOTA, JOHNS HOPKINS UNIVERSITY, Northwestern University, Georgia Institute of Technology